DOCUMENT RESUME

ED 415 687 FL 024 994

AUTHOR	Bernal, Ernesto M.
TITLE	Tests of Language Proficiency. English/Spanish Tests of Language Dominance and Proficiency.
PUB DATE	1997-01-00
NOTE	29p.; Based on a paper presented at the Annual Meeting of the Arizona Association for Bilingual Education (Phoenix, AZ, January 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Bilingual Education; Comparative Analysis; *English (Second Language); *Language Dominance; *Language Proficiency; *Language Tests; Limited English Speaking; Scoring; *Spanish; Test Interpretation
IDENTIFIERS	IDEA Oral Language Proficiency Test; Language Assessment Scales (De Avila and Duncan); Woodcock Language Proficiency Battery

ABSTRACT

Several English and Spanish language tests currently used in bilingual education and some less known tests are reviewed, looking at both historic and psychometric factors. It is suggested that practitioners have commonly assumed that all language tests are basically valid, which contributed to the commercial success of some tests of questionable validity, ill-suited to the purpose of educational decision-making. Issues discussed include the various notions of and assumptions about language proficiency and language dominance reflected in them, reliability and validity, and considerations in test selection. Appended materials include a graphic representation of how language dominance and language proficiency interact at various levels, and a brief review of three tests: the IDEA Oral Language Proficiency Tests; Language Assessment Scales; and the Woodcock Tests. Contains 26 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 415 687

Tests of Language Proficiency

RUNNING HEAD: Tests of Language Dominance and Proficiency

English/Spanish Tests of Language Dominance and Proficiency*

by

Ernesto M. Bernal

The University of Texas-Pan American

Edinburg, TX

*Based on a paper presented at the meeting of the Arizona Association for Bilingual Education, Phoenix, January 1997.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Ernesto M. Bernal

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

FL024974

Tests of Language Proficiency

Abstract

The domain of commercial language proficiency tests for children suspected of having limited proficiency in English has been particularly treacherous, especially for educators who assume that all language tests are basically good, valid instruments and that their principal task is to find, adopt, and use the one which is inexpensive and easy to administer and score. This approach, unfortunately, contributed in the past to the commercial success of language proficiency tests of poor quality, ill suited to the purposes of educational decision-making. Fortunately a number of reasonably good instruments have been developed or refined over the years to serve this purpose, although all of these require some effort to develop the skills of test administration, scoring, and interpretation.

Tests of Language Proficiency

English/Spanish Tests of Language Dominance and Proficiency

Introduction

This sampler of language dominance and proficiency tests illustrates the range of psychometric quality which these instruments represent. The domain of language testing of students for English proficiency in the past was treacherous for uninformed practitioners, who too frequently naively assumed that all language tests were basically good, valid instruments and that their principal task was to adopt and use one that was inexpensive and easy to administer and score. These attitudes, unfortunately, contributed to the initial commercial success of some language proficiency tests of questionable validity, ill-suited to the purposes of educational decision-making. Language proficiency tests help educators, such as the members of a school's Language Proficiency Assessment Committee (LPAC), to place children into bilingual/ESL programs, later to reclassify them as English-proficient students, and to determine which of the students are ready for achievement testing in English, especially for participation in statewide academic assessment programs (O'Malley & Valdez Pierce, 1994), and the types of accommodations for testing in English (such as the use of dictionaries or extra time to complete the test) that will be allowed (Olson & Goldstein, 1997;Speigel-Coleman, 1997).

The author's involvement with bilingual education since it became a federal program in 1967, and his professional acquaintance with many of the authors of past and present language tests, gives him some unique and possibly biased insights into the history of language dominance and proficiency tests' development. His having served as a member of the Texas Education Agency's Advisory Committee for the Evaluation of Language Assessment Instruments, and as an external evaluator of bilingual programs, both basic and demonstration, federal and state, has given him knowledge about the characteristics, uses, and limitations of these tests as well.

Tests of Language Proficiency

Several of the currently popular tests in bilingual education circles are reviewed in this article, and commentaries are made on a few that are less well known, since it is the purpose of this paper to review these instruments historically and psychometrically in order to provide bilingual and English-as-a-second-languge (ESL) practitioners a sense of the field of applied language proficiency testing and help them become better informed consumers.

Historical Context

Historically, many of today's language tests have their roots in the late 1960s and early 1970s (Thorum, 1981), when the nascent bilingual education movement needed language assessment instruments to identify students who should be assigned to the new programs being set up in hundreds of school districts throughout the country. Since psychometrists, counselors, and school psychologists at the time were almost exclusively English monolinguals, the responsibility for language testing devolved upon the bilingual program staff, usually teachers or aides, who were not testing specialists. These classroom personnel, furthermore, often had to conduct these tests under less-than-ideal circumstances (e.g., in a corner of their classroom) and under time constraints (e.g., testing had to be completed during the first month of the school year in time for the official count). Thus "quick and dirty" became the unspoken rule for both test-makers, who wanted to exploit the new market, and bilingual educators, whose priorities and training were, after all, more focused on teaching than on testing. Some of the more promising but relatively cumbersome instruments of the period (e.g., Gloria and David test) did not survive commercially, and today only a few of the tests (BINL, BSM, IDEA/IPT, LAB, LAS, and the Woodcock-Munoz) have sufficient scope to obtain a minimally acceptable sample of a child's language skills. Many of the rest are still in use, although some are no longer being published and survive only because of the availability of xerox technology.

on the other hand, are individually administered in whole or in part (to obtain a language sample) and require the examiner/scorer to have greater rapport-building, articulation, listening, and notational skills as well as linguistic judgment.

Clearly, tests which demand this much of the psychometrist should be examined for reliability and validity under actual field conditions. Fortunately, a few were. But most tests' technical manuals either did not report reliabilities (e.g., Dos Amigos, James, OLDM), or did not calculate reliabilities from field data. In other instances, the individual child's homeroom or self-contained classroom teacher provided the only proficiency criterion against which the test was evaluated; no attempts were made to obtain the assessments of other teachers or of competent, independent judges, such as trained linguists, to measure the construct validity of the test or at least its concurrent validity with the experts' ratings. Expediency probably motivated these oversights. Only a few of the early tests (e.g., S/ELPS) reported honest efforts to achieve this convergence.

Certain authors used "teacher judgment" as the basis for validating the language categories of levels of proficiency that their tests yield. In some instances (e.g., James) the teachers knew their students' test scores and were merely asked whether they agreed with these classifications, instead of being led to rate students' language characteristics independently of their knowledge of the results, a more valid method of measuring the congruence between scores and perceived language fluency. Other tests (e.g., IDEA/IPT, LAS, W-M) obtained independent estimates of the students' levels of proficiency in order to validate their scores and these tests have survived until the present day.

Certain technical manuals either avoided the issues of validity and reliability (Dos Amigos), made only passing or incomplete references to them (Crane), or were filled with sophisticated rationalizations and inappropriate analyses (BINL, James),

At that time, too, these tests tried to measure limited English speaking ability (LESA), not today's more inclusive constructs of limited English proficiency (LEP) and English-language learner (ELL), which involve not only aural comprehension and oral production but also the minimal reading and language arts skills appropriate for the child's grade level. All but a few of today's tests (LAB, LAS, Woodcock-Munoz) require supplementary testing of reading and writing abilities to help determine when children may exit a bilingual or ESL program with minimal risk of subsequent failure or arrested achievement.

Some tests (Crane, Dos, James) sprang full-grown from the brows of their creators during weekend fits of innovation, or so it seemed. Others were hurriedly adapted from linguistic research studies and one was actually a curriculum-bound placement test (IDEA/IPT), which was adventitiously found useful for language assessment purposes as well. Other tests (BINL, Crane, Dos, James, OLDM) were never psychometrically examined until after they were in public use--a condition which tended to make their authors and publishers somewhat selective in how they gathered and reported the data to sustain respectability and their tests' credibility.

Tests Reflect Various Notions of Proficiency and Dominance

The authors also chose to measure similar sounding constructs in a variety of ways using various scales or methods of interpreting raw scores. Some tests, for example, claimed to measure language dominance (Crane, James) without first scaling language proficiency, most tests correctly tried to measure both language proficiency and dominance by scaling English and Spanish separately and then compare the results, and others (such as the SPLIT) were designed to place students according to legal educational classifications, the Lau categories, and were actually not very useful for diagnosis or intervention.

If language proficiency is seen as scaled from 1 to 5 for both the native language

(L_1) and the second language (L_2)--English--then the results may be mapped on a 5-by-5 matrix, a Cartesian-coordinate space where both scores are plotted at one point (or, in this case, in one cell). Figure 1 depicts this matrix.

Figure 1 illustrates the fallacy of specifying language dominance without reference to proficiency, for above and below the shaded diagonal that indicates "balanced" bilingualism there exists a great range of proficiency, which could yield dominance in either L_1 or L_2, respectively. For example, a child who scores a level 5 in Spanish and a level 3 in English may be considered Spanish dominant. But this child will likely be quite different <u>as a student</u> than one who is also Spanish dominant but who scores a level 3 in Spanish and a level 1 in English. Indeed, most of the language proficiency tests do not give credit for a response which incorporates any form of language mixing or code switching, so it is not always evident by the scores alone whether a child who fails to demonstrate proficiency in either language is one who speaks a local dialect and has good communication skills, for which no credit is given according to the scoring scheme, or is a child who is not very verbal for some other reason. (See Garcia, 1979, and Lindholm & Padilla, 1978, for discussions of the cognitive and communicative implications of language interweaving.)

Nor are balanced bilingual children--those whose L_1 and L_2 scores are equal--all alike. A "5-5" child may be gifted (Bernal, 1980a), while a "1-1" or "2-2" child may require further assessment to determine if a language-related handicapping condition exists (Bernal & Tucker, 1981). The Bilingual Syntax Measure takes these differences into account, particularly at Level II (grade 3 and above). The notes at the bottom of Figure 1 indicate that a child's pair of scores (in L_1 and L_2) can be converted into hypotheses about the child that can be tested informally at first (see Guerin & Maier, 1983; Potter & Wamre, 1990), then if necessary, formally through a diagnostic evaluation. Keep in mind, too, that a child who scores a 5 in L_1 has achieved in the

normal category or perhaps in the somewhat better than average range, and therefore could not have a "language problem" other than a lack of proficiency in English. Thus the L_1 score on a well designed language proficiency test can also serve to contraindicate certain exceptionalities that would seem tenable when the child is tested exclusively in L_2 (Bernal & Tucker, 1981).

Establishing categories such as "mostly English with some Spanish" or "Spanish monolingual" are both misleading and nearly useless. The former category can encompass ten cells in Figure 1, those below the diagonal. The latter category can be used for students in any of four cells, levels 2 through 5, of the far left column. Obviously many levels of proficiency that should signal different kinds of curricular experience are subsumed under these types of classifications.

Such kinds of assignations probably rest on the notion, which still infects the thinking of many educators, that proficiency in L_1 is somehow at odds with proficiency in L_2 (Cummins, 1984). The manual for the Oral Language Proficiency Measure, for example, attempted to validate the test by correlating English and Spanish oral proficiency scores with an English-based reading achievement test. The results indicated a positive correlation between English proficiency and English reading and a negative correlation between Spanish proficiency and English reading. (Interestingly, the correlations between OLPM English and Spanish levels of proficiency decreased steadily from grades four to six, from -.43 to -.20.) Implicitly, then, the ideal was to move children who are above the diagonal in Figure 1 toward the lower right, that is, to increase their proficiency in English and to <u>replace</u> L_1 (Spanish) with L_2 (English). A better validity study would have involved the use of some current theory of bilingualism, perhaps one which sees a common cognitive basis for proficiency in either language (e.g., Cummins, 1979), to see how progressively higher levels of <u>bi</u>lingual proficiency are correlated with academic skills

such as English reading. If the relationship between English proficiency and English reading in a bilingual population is being studied, proficiency in the other language should be a covariable in the statistical solution to the design of the study.

Reliability and Validity

The reliability of tests is indicated by an index number, a correlation coefficient. In practical terms this can be seen in the size of the test's standard error of measurement (S_{EM}). Thus, any obtained score is seen as accurate within limits, for example, a score of 55 plus or minus 6. Few language proficiency test manuals actually present the S_{EM}, even if they report the test's reliability, and the reader is left either to calculate it or to ignore it.

A test's reliability is influenced by the number of items it contains. Language proficiency tests that are brief have large standard errors, so large, in fact, that their ability to assign children reliably to one or another of their own language proficiency categories is seriously in question. In short, a child who is believed to be functioning at, say, a level 3, might have only level 2 proficiency...or may be a level 4! Given that most tests only distinguish about five or six language categories or levels of proficiency, such "slippage" could be quite important both to the educator who makes placement decisions and to the child. Interestingly, the Woodcock-Munoz has intermediate categories, e.g., a level 3-4, which are particularly useful near the critical score that determines fluency in L_2, a level 4 in the case of the W-M.

Language assessment instruments are more sensitive to the efforts and skills of examiners than group tests that are commonly administered by classroom personnel. A standardized group achievement test, for example, will yield comparatively reliable results so long as the setting requirements, directions, and time limits are not violated. Students who take such tests complete their own work by marking their answers on a test booklet or a separately scorable answer sheet. Language assessment devices,

couched in impressive technical language, which probably served to discourage the reader from examining these vital technical sections carefully (Lennon, 1966) or subvert the potential customer into thinking how clever he/she would be to recommend the adoption of one of these tests for an entire district! A few publishers, however, attempted to demonstrate the educational utility of their tests (e.g, IDEA/IPT, LAS, W-M) empirically by including important correlational data in their manuals or by supplementing the original data with further studies.

Caveat Emptor, and What to Do About It

The upshot of these historical and measurement-related factors was the lack of test comparability (Ulibarri, Spencer, & Rivas, 1981). The results depended more on the test used than upon the actual language abilities of the children tested. This condition made it well-nigh impossible to recommend a language test (Oakland et al., 1980; Texas Education Agency, 1979), and the burden of validation fell entirely on the user. Few practitioners at the time realized the state of affairs, and this prompted Thorum (1981) to recommend that the decision about which test or tests to adopt be postponed until a study of some of the seemingly more promising tests would be made by administering these to a group of children in one's own district and analyzing and comparing the results.

Bernal (1997) believes that Thorum's suggestion is still a good one. Start by ordering a few specimen sets of three or four of the tests which seem attractive, have the more technically qualified and knowledgeable teachers and staff review and administer them to a representative group of LEP and non-LEP children, do a series of ministudies to see if the scores correlate with and--more importantly--predict meaningful criteria such as grades, achievement test scores, or retention in grade (a validity study), then calculate each test's reliability coefficient. Circulate the findings in the form of a detailed memorandum, solicit the opinions of people who will ultimately

use them, and finally adopt the instrument or instruments which are the most valid, reliable, and cost efficient. However, if more than one test is adopted, care must be taken to ensure that the tests yield comparable results, i.e., that the scores used for decision-making (be these standard scores, the old-fashioned Lau categories, or proficiency levels) are highly correlated.

Many districts, however, have already committed themselves to a particular language test. Once a test is in widespread use, persons responsible for educational placements would be wise to determine cutoff points empirically for different grade levels, instead of assuming that the scores or categories of the test will suffice to place students in bilingual, ESL, or regular instructional settings. Expectancy tables can be readily set up (see Wesman, 1949) to study the test's usefulness in assigning students to regular, compensatory, and special education classes as well, assuming of course that these programs represent truly different approaches suited to the language, achievement, and other cognitive characteristics of the children. These figures may be cross-validated during subsequent years, and the test or tests which emerge as reliable, predictively valid instruments may then be used with assurance to place students into (and out of) particular educational programs.

It would not be prudent to expect that a language proficiency test alone will suffice to determine students' placement or reclassification (McCollum, 1981). Educational decisions such as these require more refined instrumentation than current language proficiency tests represent, as well as broader academic content coverage to ensure that the students to be reclassified perform acceptably well in the different core academic areas. Nevertheless, a properly validated instrument should prove to be invaluable to the deliberations of a language proficiency assessment committee or other group who must determine which educational programs students will receive.

Tests of Language Proficiency

References

Bernal, E. M. (1980a). ERIC/TM Report 72: Methods of identifying gifted minority students. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service.

Bernal, E. M. (1980b). Testing Spanish native language skills in bilingual education. Paper presented at the Summer Symposium on Bilingual Multicultural Education Research, Buffalo, NY.

Bernal, E. M. (1997, January). Language proficiency testing. Paper presented at the meeting of the Arizona Association for Bilingual Education, Phoenix, AZ.

Bernal, E. M., & Tucker, J. A. (1981). A manual for screening and assessing students of limited English proficiency. Paper presented at the Council for Exceptional Children's Conference on the Exceptional Bilingual Child, New Orleans, February 1981. (ERIC Document Reproduction Service No. ED 209 785)

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. Review of Educational Research, 49, 222-251.

Cummins, J. (1984). Bilingualism and special education: Issues in assessment and pedagogy. San Diego: College-Hill.

Dalton, E. F., Amori, B. A., Ballard, W. S., & Tighe, P. L. (1991). The IDEA Oral Language Proficiency Tests. Publisher: Ballard & Tighe.

De Avila, E. A., & Duncan, S. E. (1981). A convergent approach to oral language assessment: Theoretical and technical specifications on the Language Assessment Scales (LAS) Form A. San Rafael, CA: Linguametrics Group.

De Avila, E. A., & Duncan, S.E. (1990). Language Assessment Scales. Publisher: CTB/Macmillan-McGraw-Hill.

Garcia, E. E. (1979). Bilingualism: A developmental psycholinguistic approach in K. A. Martinez & S. Arenas (Eds.) Bilingual/multicultural early childhood education:

Proceedings of Head Start regional conference, 1978-79. Washington, D.C.: Administration for Children, Youth and Families.

Guerin, G. R., & Maier, A. S. (1983). Informal assessment in education. Palo Alto, CA: Mayfield.

Lennon, R. T. (1966). The test manual as a medium of communication. In A. Anastasi (Ed.), Testing problems in perspective. Washington, D.C.: American Council on Education.

Lindholm, K. J., & Padilla, A. M. (1978). Child bilingualism: Report on language mixing, switching, and translations. Linguistics, 211, 23-44.

McCollum, P. A. (1981). Concepts in bilingualism and their relationship to language assessment. In J. G Erickson & D. R. Omark (Eds.), Communication Assessment of the Bilingual Child. Baltimore, MD: University Park Press.

Oakland, T., Bernal, E. M., Holley, F., et al. (1980, September). Assessing students with limited English speaking abilities. Texas Outlook, pp. 32-33.

Olson, J. F., & Goldstein, A. A. (1997). The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress (NCES R & D Report 97-482). Washington, DC: U. S. Department of Education, Office of Educational Research and Improvement.

O'Malley, J. M., & Valdez Pierce, L. (1994). State assessment policies, practices, and language minority students. Educational Assessment, 2 (3), 213-255.

Potter, M. L., & Wamre, H. M. (1990). Curriculum-based measurement and developmental reading models: Opportunities for cross-validation. Exceptional Children, 57 (1),16-25.

Speigel-Coleman, S. (1997, August). At what point are LEP children ready to be tested in English? Approaches to determining readiness. Paper presented at the High Stakes Assessment for LEP Students Conference, U. S. Department of Education,

Office of Bilingual Education and Minority Languages Affairs, Washington, DC.

Texas Education Agency. (1979). Report of the Committee for the Evaluation of Language Assessment Instruments. Austin, TX: TEA.

Thorum, A. R. (1981). Language Assessment instruments. Springfield, IL: Charles C. Thomas.

Ulibarri, D. M., Spencer, M. L. & Rivas, G. A. (1981, Spring). Language proficiency and academic achievement. Relationship to school ratings or predictions of academic achievement. NABE Journal, 5 (3), pp. 47-80.

Wesman, A. G. (1949). Expectancy Tables--a way of interpreting test validity (Test Service Bulletin, No. 38). New York: Psychological Corporation.

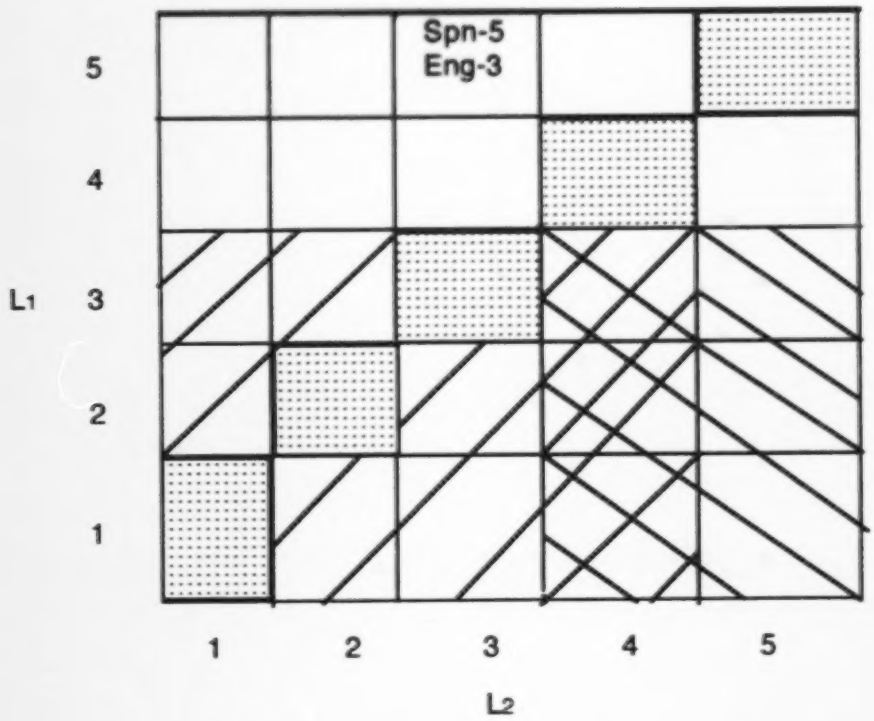
Woodcock, R. W. (1991). Woodcock Language Proficiency Battery-Revised-English. Chicago: Riverside.

Woodcock, R. W., & Munoz-Sandoval, A. F. (1993). Woodcock-Munoz Language Survey. Chicago: Riverside.

Woodcock, R. W., & Munoz-Sandoval, A. F. (1995). Woodcock Language Proficiency Battery-Revised-Spanish. Chicago: Riverside.

Tests of Language Proficiency

Figure 1. Language Proficiency and Dominance Matrix



Notes: Shaded areas indicate points of balanced proficiency, e.g., 4-4.

indicates children whose communicative competency needs to be checked to see if low scores on both L_1 and L_2 were an artifact of the test or whether a langauge-related handicap is a possibility.

indicates children who may be experiencing "subtractive" bilingualism, who are losing L_1 as they acquire L_2.

suggests that these children may have been transitioned out of L_1 permaturely and are now struggling with their acquisition of L_2.

Tests of Language Proficiency

A Brief Review of Three Language Proficiency Tests

I. The IDEA Oral Language Proficiency Tests (IPT). Publisher: Ballard & Tighe. Principal author: Dr. E. F. Dalton. Other authors: B. A. Amori, W. S. Ballard, and P. L. Tighe. All IPT tests are administered individually in approximately 15 minutes.

IPT I - English - Forms C & D, grades K to 6, 1991.

Forms C & D are intended to replace the 1979 Forms A & B. "Field studies" were conducted in 21 LEAs in seven states. The instrument tests students' syntax, lexicon, comprehension, morphology, and oral expression (including articulation). It can be scored to yield NES, LES, and FES designations as well as six levels of proficiency, A to F (F meaning fluent), which are described in terms of oral skills and are ostensibly linked to diagnostic/placement/ instructional levels of the IDEA kit, from which the original forms of the IPT sprang. Students at different grades (or ages, it is presumed) usually begin at different levels to accommodate their linguistic growth. Instructions for the test are given in the child's primary language, but prompting during the test is done entirely in the target language of the test. Each level has a STOP score to terminate the testing, and credit is given for previous levels achieved. These achieved levels can be converted to the Lau designations; at grade 1, for instance, Levels A or B = NES, Levels C or D = LES, and Levels E or F = FES.

In addition to content validity, the authors claim construct validity because "the student is led through progressively more difficult items" in the content covered. A contingency coefficient based on an 8 x 6 Chi-square (ages 5 to 12 x Levels A to F) for Form C was calculated at .37 and for Form D at .40, indicating a positive correlation between age and level, but the Pearson r, which is calculated on score, would be the more appropriate statistic. The rs for Forms C and D are, respectively, only .30 and .27.

Very interestingly, "English-only" students were used to corroborate the English test, so that ELLs would not be called upon to use forms that native speakers at these

ages have not mastered. Content validity and instructional utility with the IDEA curriculum seem to have been established.

A better construct validity study would have correlated scores (not levels) on the different aspects of language tested, such as syntax and morphology and comprehension, with age and other personal indicators of the students, such as time in the U. S. Similarly, concurrent validity, studied by correlating teachers' estimates of IPT Levels with obtained IPT levels, could have been better corroborated at the higher Levels (E/F) or the highest designation (FES) with language-related achievement tests in English, since Levels and designations are used for placement decisions. School districts' own designations were correlated with the IPT's Levels and yielded r_s of .65 and .66 for Forms C and D. But this study should have been made using the proficiency designations from the IPT, and this would have yielded a measure of categorical/diagnostic validity.

Reported inter-rater reliabilities for Forms C and D are .93 and .87 based on levels, not by items, and so the reported alpha coefficients are astronomically high, .99 for both forms. The test's statistics, while reported in some detail, were computed under most favorable circumstances, e.g., by correlating Levels, which cover a range of scores. Clearly the IPT tests have a higher standard error than the reported data imply. Because of this what appears to be an otherwise good test can be recommended only with reservations.

Finally, while the Group List provides for entering the proficiency levels and designations of students for both languages (English and Spanish), no special curricular or placement recommendations arise from the use of these pieces of information taken together. Bernal (1980b) and Bernal and Tucker (1981) earlier demonstrated that proficiency levels considered simultaneously can assist in the initial screening for real language difficulties (not just low achievement) and even help

nominate candidates for gifted program assessment.

The IPT-I-English is the latest and probably the best prepared of the IDEA tests.

IPT I-Spanish-grades K to 6, 1989.

This test is very similar in rationale and composition to the IPT-I-English, but has a much smaller field-trial sample and older data (1980), which were collected when the test covered grades K-8. For this revision an additional K-level sample of 168 students was added in 1988 to validate the NSS/LSS/FSS designations at that level.

IPT II-English-Forms A & B-grades 7 to 12, 1983.

This test replaces the 1979 edition. Used 306 "English-only" students to validate the items, but only 153 students to test the NES/LES/FES designations. Between-forms reliability = .75 when levels are used, and canonical correlation between teachers' NES/LES/FES designations and the obtained designations from the test was only .45; discriminant analysis placed only 51.6 percent (Form A) and 61.5 percent (Form B) of the students in the same designations. Clearly this test is not sufficiently refined for use with ELLs for purposes of reclassification.

IPT II-Spanish-grades 7 to 12, 1987.

Like all of the tests in this series, Levels A to F represent similar skills in both English and Spanish versions. This particular test, however, claims to measure cognitive/academic language proficiency (CALP). This test underwent field trials with an N of 481 in Texas and California. Test-retest reliability was only .62. The authors report r_s between age/grade and IPT Levels in an attempt to establish construct validity, and obtained very low correlations; it is doubtful if at these ages such correlations have anything to do with construct validity. The low test-retest reliability, computed under the most favorable conditions by correlating Levels (not raw scores), suggests that the IPT II-Spanish-Grades 7 to 12 is inadequate for educational placement purposes.

Tests of Language Proficiency

Summary of IDEA/IPT Tests.

The better IDEA/IPT tests are those used for grades K to 6, which are the grades where the greater need is to be found. The designations for NES/LES and NSS/LSS are probably the most treacherous, since these do not seem to be based on validatable expert opinion or objective performance criteria. FES and FSS are probably sufficient to recognize the more proficient speakers at the highest level or two, but there is no evidence that reclassifications based on these designations or Levels produce satisfactory results. Indeed, one of the shortcomings of the test reveals its close ties to the IDEA curriculum: Several school district psychometrists attending the Arizona Association for Bilingual Education convention in 1997 complained that students transitioning between IPT levels show a sudden drop in proficiency. In short, there are no predictive validity studies based upon followup of reclassified students.

Conclusion: IDEA/IPT Tests.

Use with caution; better yet, do your own validity study before endorsing these instruments for any use other than placement in the IDEA curriculum, which is what the tests were initially designed to do and in all liklihood do best.

II. Language Assessment Scales (LAS). Publisher: CTB/Macmillan-McGraw-Hill. Authors: E. A. De Avila, Ph.D. and S. E. Duncan, Ph.D. All LAS tests are individually administered in approximately 20 to 25 minutes. The LAS tests include cassettes and picture booklets and require audio taping of the Story Retelling sections. Scoring services are available. Separate scoring/interpretation and administration manuals are provided.

The LAS test has been used for 20 years. The latest revisions of these instruments (1990) are organized as follows:

LAS-Oral (LAS-O)

- PRE-LAS (grades Pre-K to 1) Forms A & B

Tests of Language Proficiency

- LAS-O (grades 1 to 6) Forms 1C & 1D - English
- LAS-O (grades 1 to 6) Form 1B - Spanish
- LAS-O (grades 7 to 12) Forms 2C & 2D - English
- LAS-O (grades 7 to 12) Form 2B - Spanish
- LAS-O Pronunciation (grades 2 to 12) Forms C & D

The "long" forms of the LAS-O contain the Pronunciation component.

-LAS-Reading & Writing (LAS-RW)

Several of the LAS tests are briefly reviewed below, and comments about the entire battery follow.

LAS-O Form 1B - Spanish, grades 1 to 6.

The parts of the test include vocabulary, listening, comprehension (stories), and the pronunciation (optional) component, which consists of minimal pairs (phoneme distinction) and enunciation. The administration manual includes directions for administration in both English and Spanish.

LAS-O Forms 1C and 1D - English, grades 1 to 6.

These parallel forms replace the old Forms A and B. Format parallels the Spanish test at Level 1, but the content is original to the English version. Cross-tabulations of Forms 1C and 1D levels yield a Pearson r of .91, a 77 percent agreement.

LAS-O Form 2B - Spanish, grades 7 to 12.

This test replaces the old Form A. The format parallels the format of the other LAS tests, but the Spanish content at this level is original.

LAS-O Form 2C and 2D - English, grades 7 to 12.

These forms replace the older Forms A and B. Parallel form correlation = .96 for the long forms. Cross tabulations of levels of proficiency for English Forms 2C and 2D yield an r of .92 and an agreement of 79 percent.

Tests of Language Proficiency

Summary of LAS Tests.

The LAS tests are probably the most studied language proficiency tests on the market, and the authors have taken few shortcuts to ensure the adequacy of the test. For certain levels of the test, exercises are provided to assist users in training those who score the Story Retelling part of the test, for example. In addition to fairly standard types of psychometric analyses, the LAS reports the development of special scales, such as the Language Proficiency Index, which attempts to combine aspects of literacy (reading, writing) with oral proficiency into one score that might be useful for reclassification decisions, for instance.

The "tryout" sample for the English tests consisted of 3,600 elementary and secondary students from nine LEAs in five states. Of these students 1,172 had English as their home language. Subscale reliabilities (r_{aa}) are mostly in the high .80s and distinguish reliably between fluent and limited speakers. Inter-rater reliabilities vary from the .30s and .40s for Listening Comprehension to the high .80s and low .90s for the other subscales.

The Spanish tests were validated against 1,264 students in three sites. Alpha (r_{aa}) coefficients for subscales vary from .77 (Listening Comprehension, Form 2B) to .96 (Vocabulary, Form 1B). Native speakers assured that the items selected were realistic and helped to establish the important distinction between a child who has normal development in L_1 and merely lacks proficiency in L_2, and the child who is weak in both.

Conclusion: LAS Tests

The English tests are the ones that have been most extensively revised and the ones which will be most useful in selecting and placing students initially into special programs for ELLs. But the LAS should be particularly useful for reclassifying students, especially when the reading/writing exams supplement the oral tests to

ensure the development of sufficient academic skills in English to succeed in school outside of the sheltered language environment of a bilingual or ESL classroom.

III. The Woodcock Tests.

To understand the composition and function of the Woodcock-Munoz Language Survey (W-M) in English and Spanish, it is necessary to see the W-M in relation to the Woodcock Language Proficiency Battery-Revised (WLPB-R) and the Woodcock Language Proficiency Battery-Revised-Spanish (WLPB-R-S) from which it was derived.

A. Language Proficiency Battery-Revised-English (WLPB-R), 1991. Publisher: Riverside. Author: R. Woodcock, Ph.D.

This test replaces the original (1984) version by adding five new subtests (for a total of 13 subtests) and extending the norms downward to 24 months and upward to adults over 50 years of age. The test is administered through the use of an easel test book and one audio tape for presenting two of the subtests. True to the Woodcock tradition, the WLPB-R covers several language areas so that the results may have diagnostic value and assist in individual educational program planning: Oral Language (tests 1 to 5), Reading (tests 6 to 9), and Written Language (tests 10-13).

1. Memory for Sentences: repeating phrases and sentences
2. Picture Vocabulary: naming familiar and unfamiliar pictured objects
3. Oral Vocabulary: synonyms and antonyms
4. Listening Comprehension: an oral cloze procedure
5. Verbal Analogies: complete a logical work relationship
6. Letter-Word Identification
7. Passage Comprehension
8. Word Attack
9. Reading Vocabulary: synonyms and antonyms

10. Dictation

11. Writing Samples: quality of expression

12. Proofing: identification of mistakes

13. Writing Fluency: writing single sentences in response to a stimulus-picture.

Spelling, Usage, and Handwriting may also be checked.

These subtests can be combined to yield scores in Broad English Ability (Early Development), Broad English Ability (Standard Scale), an Oral Language Cluster (tests 1 to 5), three Reading clusters, and three Writing clusters. Clearly the WLPB-R contains contextually embedded as well as contextually reduced components (see Cummins, 1984) and therefore presents diverse cognitive challenges that go beyond simple definitions of proficiency. Testing time for the entire battery requires approximately one-and-a-half hours. To facilitate this process, basal and ceiling rules of testing apply, which make testing more time-efficient and not especially frustrating for students at the upper limits of their achievement.

Scoring on 11 of the 13 tests is straightforward: an item earns either a 1 or a 0. Directions for test administration are very clearly written. Raw scores are converted to a variety of scales complete with confidence intervals, some on the protocol itself. Special scores, such as Woodcock's adaptation of the Rasch ability scores (W-scores, anchored at 5th grade, with a mean score of 500) and a criterion-referenced Relative Mastery Index (RMI) can also be calculated. Percentiles and standard scores (the latter based on a mean of 100 and a sigma of 15) can be found in a separate book of tables. Practice in scoring for would-be test administrators is also provided in the manual.

Internal consistency measures range from a low of .66 (Listening Comprehension at age 13) to a high of .98 (Letter-Word Identification at age 70-79, and Writing Samples at age 6), but most are in the .80s and .90s. Cluster reliabilities

are higher, on average. Inter-rater reliabilities are also reported for not too sophisticated scorers and range from .75 to .99. Very importantly, test-retest reliabilities corrected for age range from .75 to about .90--very respectable indeed.

Content validity and concurrent validity studies are discussed.

B. Woodcock Language Proficiency Battery-Revised-Spanish (WLPB-R-S), 1995. Publisher: Riverside. Authors: R. Woodcock, Ph.D., & A. Munoz-Sandoval, Ph.D.

"All tests in the Spanish WLPB-R have been adapted from the parallel English Form..." (Woodcock & Munoz-Sandoval, 1995, p.1). Reference is specifically made by the authors to two interpretive features: (1) Cummins' cognitive-academic language proficiency (CALP), and (2) the Comparative Language Index (CLI) "that allows direct comparison of Spanish and English language proficiencies in a single index..." (p.1). The norms for the English form obtain throughout, although Spanish items were equated to English items using the Rasch model on some 2,000 Spanish monolinguals in the United States and in several countries in Latin America. The calibration to the English version was based upon a subset of English items, representing a wide range of difficulty, that were translated into Spanish and administered as an "anchor" test. These items were developed using the back-translation method for checking equivalent meanings.

Five levels of CALP in English and Spanish are defined according to the W-score differences and the RMIs for a student's attained age:

Level 5: Advanced CALP

Level 4: Fluent CALP

Level 3: Limited CALP: finds the language demands of monolingual instruction in this language difficult.

Level 2: Very Limited CALP

Level 1: Negligible CALP.

Intermediate levels based on the SEM of these levels (6 W-units) are also proposed, and are particularly useful in the greyhh area between levels 3 and 4, where crucial placement decisions are most often made.

The Comparative Language Index (CLI) is derived from the Oral Language Clusters of the English and Spanish tests. Specifically, the Spanish Relative Proficiency Index (RPI) and the English Relative Mastery Index (RMI) are compared for an individual student in a S/E ratio, which is keyed in turn to the five CALP levels.

Measures of internal consistency for individual tests of the Woodcock-Munoz range from .51 (Correccion de Textos for age 6) to .98 (Vocabulario Sobre Dibujos and Identificacion de Letras y Palabras for age 70-79), but are generally in the mid-.80s for children of school-age.

The authors' interpretation of CALP appears to be a good one, generally, but the distinction between CALP levels and raw proficiency levels needs to be kept in mind. Level 1 CALP could be a Level 2 speaker of the language in terms of Figure 1, indicating a student who is not a raw beginner but who clearly cannot yet think in the language being tested. Recall that the five-level system was initially developed for use with adults learning a foreign language, for whom the acquisition of CALP was never an issue.

The decision to base the Spanish version entirely on the responses of Spanish monolinguals may help make the Woodcock-Munoz a commercial success throughout Latin America. The Woodcock-Munoz should be an excellent test to use with recent Spanish-speaking immigrants. Psychometrically this also makes sense because unrealistic language expectations can be eliminated.

However, there is also the possibity that the resultant test may systematically underestimate the total communicative competence of native-born U. S. bilinguals who, not ordinarily being educated through the medium of Spanish nor exposed to a

lot of higher-register English, may switch language codes fluently, i.e., may speak "Tex-Mex" as their native tongue. If mixed-language responses are acceptable in either the English or the Spanish tests or both, the directions for scoring are not explicit about this problem. If mixed-language responses are acceptable in either or both tests, does the Comparative Language Index not then become useless? This author fears that the Woodcock-Munoz, for all its academic sophistication and diversity of content, could go the way of other, earlier language tests and yield suspiciously low scores on both English and Spanish for many Hispanic students in the United States, reinforcing the notion that "these students don't speak either language well" and intimating that they have a developmental language disability, when in fact they speak Tex-Mex or some other local or regional dialect very well and have excellent communicative competence among their peers. Bilingual educators and advocates of bilingual education should watch for such untoward developments, which could have bearing on the consquential validity of the Woodcock-Munoz test. One may wish to test a number of known normal and linguistically delayed Tex-Mex speakers, five or six years of age, to see if the Woodcock-Munoz can tell the difference.

It is also possible that the order of testing can affect the outcomes. One might argue about the sequence, but I recommend that testing first be done in what the examiner believes is the student's <u>weaker</u> language, based upon the language-use questions at the beginning of the test protocol. One may assume that the student tested in the <u>stronger</u> language first will reach items beyond the ceilings that would be encountered in his/her weaker language, and that this familiarity with the test content at these levels could contaminate the results, making the student's performance in the weaker language appear stronger than it is, with the result that English-language learners are initially placed into non-bilingual, non-ESL classrooms when in fact they require these specialized educational interventions.

Tests of Language Proficiency

C. Woodcock-Munoz Language Survey (W-M), 1993. Publisher: Riverside. Authors: R. Woodcock, Ph.D., & A. Munoz-Sandoval, Ph.D.

This is the version of the Woodcock language tests typically used for language proficiency assessment, the version that can the administered in 15 to 20 minutes per language to most students, the one known simply as the "Woodcock-Munoz." The W-M consists of four tests taken from the WLPB-R and four tests taken from the WLPB-R-S: (1) Picture Vocabulary/Vocabulario Sobre Dibujos, (2) Verbal Analogies/Analogias Verbales, (3) Letter-word Identification/Identificacion de Letras y Palabras, and (4) Dictation/Dictado. These tests allow an estimate of the Broad English/Spanish Ability Cluster and estimates of the more specific Oral Language Clusters and the Reading-Writing Clusters.

For a more detailed view of a student's language skills, then, the socres on the W-M can be transferred to the WLPB-R or the WLPB-R-S and the remaining subtests of these instruments can be administered in standard fashion to secured the diagnostic benefits of the complete batteries.

A brief language-use questionnaire is included on the test protocol that could be useful in guiding close placement decisions. Allowable accommodations for exceptional children are discussed, and scoring and training procedures are detailed in the Comprehensiv e Manual of the test. A Scoring and Reporting software program is also available for data recording, and it yields standard screening conclusions in printed form.

Conclusion: The Woodcock-Munoz Language Survey (W-M).

The W-M assesses both receptive and expressive levels of language proficiency in English and Spanish. It is also linked to established tests of language-related achievement, which augurs well for its predictive validity in placement-exit decisions. For students who score in the "grey" area between CALP levels 3 and 4 in English, the

Tests of Language Proficiency

Woodcock-Munoz can segue into the more complete language assessment battery in English, the WLPB-R, which can shed more light on their levels of achievement and their readiness to engage the all-English classroom.